Exquisitor: Breaking the Interaction Barrier for Exploration of 100 Million Images

Hanna Ragnarsdóttir
Reykjavik University
Reykjavik, Iceland
hannar15@ru.is

Þórhildur Þorleiksdóttir
Reykjavik University
Reykjavik, Iceland
thorhildurt15@ru.is

Omar Shahbaz Khan
IT University of Copenhagen
Copenhagen, Denmark
omsh@itu.dk

Björn Þór Jónsson
IT University of Copenhagen
Copenhagen, Denmark
bjorn@itu.dk

Gylfi Þór Guðmundsson
Reykjavik University
Reykjavik, Iceland
gylfig@ru.is

Jan Zahálka
bohem.ai
Prague, Czech Republic
jan.zahalka@bohem.ai

Stevan Rudinac
University of Amsterdam
Amsterdam, Netherlands
s.rudinac@uva.nl

Laurent Amsaleg
CNRS–IRISA
Rennes, France
laurent.amsaleg@irisa.fr

Marcel Worring
University of Amsterdam
Amsterdam, Netherlands
m.worring@uva.nl

ABSTRACT

In this demonstration, we present Exquisitor, a media explorer capable of learning user preferences in real-time during interactions with the 99.2 million images of YFCC100M. Exquisitor owes its efficiency to innovations in data representation, compression, and indexing. Exquisitor can complete each interaction round, including learning preferences and presenting the most relevant results, in less than 30 ms using only a single CPU core and modest RAM. In short, Exquisitor can bring large-scale interactive learning to standard desktops and laptops, and even high-end mobile devices.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; **Multimedia databases**;

KEYWORDS

Interactive multimodal learning; Scalability; 100 million images.

ACM Reference Format:
Hanna Ragnarsdóttir, Þórhildur Þorleiksdóttir, Omar Shahbaz Khan, Björn Þór Jónsson, Gylfi Þór Guðmundsson, Jan Zahálka, Stevan Rudinac, Laurent Amsaleg, and Marcel Worring. 2019. Exquisitor: Breaking the Interaction Barrier for Exploration of 100 Million Images. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3343031.3350580

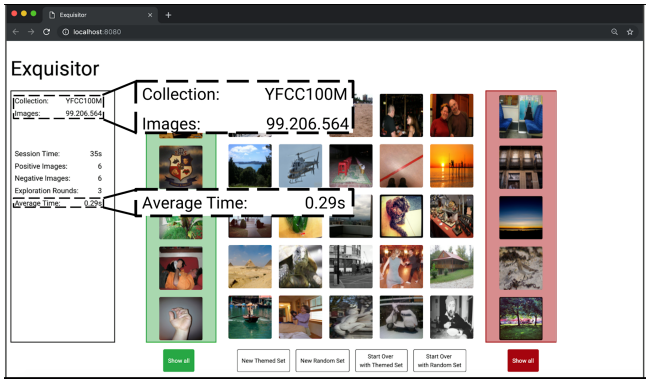

Figure 1: The Exquisitor demonstration interface. Exquisitor interactively learns new analytic categories over the full YFCC100M image collection, with average latency of less than 30 ms per interaction round, using hardware comparable with standard desktops and modern mobile devices.

1 INTRODUCTION

Multimedia collections have become a cornerstone data resource in a variety of scientific and industrial fields. One of the most difficult challenges for interactive exploration of such collections—not only for data scientists working with these collections, but also for the multimedia community as a whole—is their *scale.* How can we facilitate efficient access to multimedia collections comprising tens or hundreds of millions of images, let alone billions?

Interactive learning, which was embraced by the multimedia community in the early days of content-based image and video retrieval [4, 12], has recently experienced revival as an umbrella method capable of satisfying a variety of multimedia information needs, ranging from exploratory browsing to seeking a particular known item [17]. Through feedback from the user, interactive learning can adapt to the intent and knowledge of the user, and thus *collaborate* with the user towards, e.g., learning new or unknown analytic categories on the fly. Previous contributions, however, largely operated at a relatively small scale [1, 2, 8, 10, 13, 14].

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10.
https://doi.org/10.1145/3343031.3350580

Consider, as an example, the YFCC100M collection [16], comprising 99.2M images and 0.8M videos. It has existed for some time now, but very few have a good idea about what its actual contents are. This is no surprise, as it is difficult to tackle this collection with existing techniques: the simple metadata-based filtering approach is impeded by the sparse and noisy nature of the metadata and, at this scale, similarity search is akin to shooting in the dark. Semantic concept detectors can be used to generate additional content-based metadata for both approaches, but that does not alleviate their problems. And thus, the YFCC100M collection remains a mystery.

We have recently developed Exquisitor, a highly scalable interactive multimodal learning approach [6]. A key feature that sets Exquisitor apart from related approaches is its scalability: Exquisitor can retrieve suggestions from 100 million images with sub-second latency, using extremely modest computing resources, thus breaking the interaction barrier for large-scale interactive learning. In this demonstration, we propose to allow ACM Multimedia attendees to interactively explore the YFCC100M collection with Exquisitor.

2 EXQUISITOR INTERFACE

The current Exquisitor user interface, shown in Figure 1, is browser-based and implemented using the React JavaScript library. It is a fairly traditional interactive learning interface, in that users are asked to label positive and negative examples, which are then used to learn their preferences and determine the new round of suggestions. Due to the extreme efficiency of the interactive learning process, however, there are some notable differences from traditional interactive learning interfaces:

- The learning process runs unobtrusively in the background, continuously providing new on-demand relevant examples as the user progresses with her exploration, instead of requiring explicit management.
- Individual images are replaced, rather than the entire screen, for a smooth transition from one interaction round to the other. Users are allowed to indicate that images are neither positive nor negative to get a new suggestion, and images that have been visible for some time, but not tagged as positive or negative, can also be replaced with new suggestions.
- Users can revisit positive and negative examples, removing or even reversing the feedback label, as their understanding of the collection contents and its relevance evolves.

Overall, the user interface is intended to provide a smooth learning experience. We have already used Exquisitor in the Lifelog Search Challenge (LSC) 2019 [7], and a detailed evaluation of the user experience is part of our future work.

3 THE LEARNING PROCESS

Exquisitor's back-end produces relevant results to show to the user in less than 30 ms per interaction round, including learning the user preference and scoring the collection, using one 2.4 GHz CPU core and less than 6 GB of RAM. The back-end system is composed of two web services, as shown in Figure 2. The ImageAPI service serves thumbnails for the YFCC100M image collection, as required by the user interface. The LearningAPI service wraps the underlying multimodal learning engine, decribed in [6]. In the remainder of this section, we briefly outline the multimodal learning process.

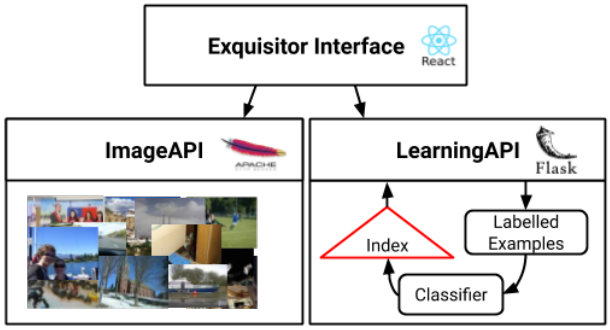

Figure 2: Exquisitor demonstration system overview.

To enable interactive multimodal learning, visual and text features were extracted from the images of the YFCC100M collection. For visual features, the 1000 ImageNet semantic concepts were extracted using a GoogLeNet architecture [15]. For the text modality, 100 LDA topics were extracted from the image title, tags and description using the gensim framework [11].

Uncompressed features for the 99.2M images require nearly 880GB of memory. Exquisitor uses the recently proposed Ratio-64 representation, which preserves only the top visual concepts and text topics for each image [18]. The compressed feature collections require less than 6GB of storage, thus fitting into the memory of a standard consumer PC, as well as some high-end mobile devices. This effective compression method has been shown to preserve the semantic descriptiveness of the visual and text features [18].

The interactive learning process is facilitated using a linear SVM model, proven to provide a good balance between efficiency and accuracy when classifying large datasets based on few training examples [5, 9]. Based on the relevance indication provided by the user, a classifier is trained separately for the text and visual modalities and the images furthest from the hyperplane are selected. The final list of results is created using rank aggregation.

The compressed feature data is indexed using a variant of the extended Cluster Pruning (eCP) high-dimensional indexing algorithm [3]. By directing Exquisitor's attention to the clusters with representatives most relevant to the learned linear SVM model, the work of scoring candidates is reduced by nearly two orders of magnitude, with an actual increase in quality [6]. The combination of all these state-of-the-art methods enables an interaction round of less than 30 ms on average using only limited computational resources: one 2.4 GHz CPU core and less than 6 GB of RAM.

4 DEMONSTRATION

The main emphasis of the demonstration will be to allow conference participants to explore the 99.2 million images of the YFCC100M collection using Exquisitor. The authors will also prepare some interesting exploration scenarios that highlight aspects of the YFCC100M collection. During the demonstration we hope to engage conference participants in a discussion that can inspire the multimedia community to work on scalable multimedia techniques and applications.

REFERENCES

[1] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. 2018. The Power of Ensembles for Active Learning in Image Classification. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Salt Lake City, UT, USA, 9368–9377.

[2] Kashyap Chitta, Jose M. Alvarez, and Adam Lesnikowski. 2018. Large-Scale Visual Active Learning with Deep Probabilistic Ensembles. arXiv:1811.03575. (2018), 10 pages.

[3] Gylfi Þór Guðmundsson, Björn Þór Jónsson, and Laurent Amsaleg. 2010. A Large-scale Performance Study of Cluster-based High-dimensional Indexing. In *Proc. International Workshop on Very-large-scale Multimedia Corpus, Mining and Retrieval (VLS-MCMR)*. ACM, Firenze, Italy, 31–36.

[4] Thomas S. Huang, Charlie K. Dagli, Shyamsundar Rajaram, Edward Y. Chang, Michael I. Mandel, Graham E. Poliner, and Daniel P. W. Ellis. 2008. Active Learning for Interactive Multimedia Retrieval. *Proc. IEEE* 96, 4 (2008), 648–667.

[5] Prateek Jain, Sudheendra Vijayanarasimhan, and Kristen Grauman. 2010. Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Active Learning. In *Proc. Conference on Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., Vancouver, BC, Canada, 928–936.

[6] Björn Þór Jónsson, Omar Shahbaz Khan, Hanna Ragnarsdóttir, Þórhildur Þorleiksdóttir, Jan Zahálka, Stevan Rudinac, Gylfi Þór Guðmundsson, Laurent Amsaleg, and Marcel Worring. 2019. Exquisitor: Interactive Learning at Large. arXiv:1904.08689. (2019), 10 pages.

[7] Omar Shahbaz Khan, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2019. Exquisitor at the Lifelog Search Challenge 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2019*. ACM, Ottawa, ON, Canada, 7–11.

[8] Akshay Mehra, Jihun Hamm, and Mikhail Belkin. 2018. Fast Interactive Image Retrieval using Large-Scale Unlabeled Data. arXiv:1802.04204. (2018), 15 pages.

[9] Ionuț Mironică, Bogdan Ionescu, Jasper Uijlings, and Nicu Sebe. 2016. Fisher Kernel Temporal Variation-based Relevance Feedback for Video Retrieval. *Computer Vision and Image Understanding* 143 (2016), 38 – 51.

[10] Karl Ni, Roger A. Pearce, Kofi Boakye, Brian Van Essen, Damian Borth, Barry Chen, and Eric X. Wang. 2015. Large-Scale Deep Learning on the YFCC100M Dataset. arXiv:1502.03409. (2015), 5 pages.

[11] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[12] Yong Rui, Thomas S. Huang, and Sharad Mehrotra. 1997. Content-Based Image Retrieval with Relevance Feedback in MARS. In *Proc. International Conference on Image Processing (ICIP)*. IEEE Computer Society, Santa Barbara, CA, USA, 815–818.

[13] Nicolae Suditu and François Fleuret. 2012. Iterative relevance feedback with adaptive exploration/exploitation trade-off. In *Proc. ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, Maui, HI, USA, 1323–1331.

[14] Nicolae Suditu and François Fleuret. 2016. Adaptive Relevance Feedback for Large-Scale Image Retrieval. *Multimedia Tools Appl.* 75, 12 (2016), 6777–6807.

[15] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going Deeper with Convolutions. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Boston, MA, USA, 1–9.

[16] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research. *Commun. ACM* 59, 2 (2016), 64–73.

[17] Jan Zahálka and Marcel Worring. 2014. Towards Interactive, Intelligent, and Integrated Multimedia Analytics. In *Proc. IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE Computer Society, Paris, France, 3–12.

[18] Jan Zahálka, Stevan Rudinac, Björn Þór Jónsson, Dennis C. Koelma, and Marcel Worring. 2018. Blackthorn: Large-Scale Interactive Multimodal Learning. *IEEE Transactions on Multimedia* 20, 3 (2018), 687–698.